whole genome duplication
whole genome duplication

both genes

one gene
whole genome duplication

both genes

one gene
whole genome duplication
gene-level relationships
gene-level relationships
gene-level relationships
gene-level relationships
gene-level relationships
gene-level relationships
contributions

**Pathline**
*first interactive tool for visualizing multiple genes, time points, species, and pathways*

**curvemap**
*visual encoding of temporal gene expression*

**linearized pathway**
*representation for comparing quantitative data along a pathway*

**validation**
*case studies describing efficiency gains and new biological findings*

a methodological approach to visualization development makes effective design decisions salient
processes in which end-users influence how a design takes shape
process

1. understand*
2. abstract*
3. design*
4. implement
5. validate* *receive feedback from scientists and refine
process

1. understand*

2. abstract*

3. design*

4. implement

5. validate*

*receive feedback from scientists and refine
abstract
data
**metabolic pathways**
- 10 to 50 pathways of interest
- inputs/outputs called metabolites

**gene expression**
- 6000 genes and 140 metabolites
- 6 time points
- 14 species of yeast

**similarity scores**
- aggregate time series for a gene/metabolite over species
- similarity of expression across species
- aggregate: Pearson, Spearman, others

**phylogeny**
- tree diagram showing relationships among species and time points
tasks
1. study expression data as a time series
   • look for peaks, valleys, time shifts

2. detailed comparison of a limited number of time series
   • filter using pathways
     • filter again using genes or species

3. comparison of similarity scores of genes along a pathway(s)

4. multiple similarity score comparison
design
validate
validation

- is the abstraction right?
- does the design support the data and tasks?

A Nested Process Model for Visualization Design and Validation
T. Munzner, IEEE InfoVis 2009.

- can the tool provide:
  - efficiency gains?
  - new insights?
The biologists verified that Pathline can show known information more clearly than could be seen with their previous tools, and they directly attribute new insights into their data to the use of Pathline.

9. Case Studies

We collaborated with a team of seven biologists who have been collecting and analyzing comparative functional genomics data for several years — one of the authors on this paper is a member of the team. We conducted weekly meetings over the course of three months with members of the team to learn about their scientific questions, analysis needs, and available visualization tools. The team was previously using conventional heatmaps generated using Java TreeView selects a single pathway, which the system shows side by side with the time series data for the genes contained within the pathway. The time series data are shown using a small multiples matrix where each axis is an experimental condition, and each grid cell has a parallel coordinates view of the gene set over the time points. Any time point in a parallel coordinates view can be expanded to show a bar chart of the expression level of all the genes. Extending this approach to the time series curves that we show would disrupt the shape perception required for many of the detailed analysis tasks of our collaborators. As we discuss in Section 6.1, seeing only a single pathway at once is a limitation of GeneShelf that we explicitly address in Pathline, as is the problem of encoding aggregate quantitative values directly along the pathway.

9.2. Whole Genome Duplication

A whole genome duplication event occurred in yeast some 150 million years ago. An ancestral yeast species gained an extra copy of all its genes, and thus living descendants of that common ancestor often have multiple copies of genes as a result of that duplication. Two such genes are g1 and g2, which are shown in Figure 4. Scientists know that the genetic controls responsible for the activity levels of these genes have evolved to behave differently. A telltale sign of this phenomenon is the shift in the peaks of the time series curves for the g2 gene activity patterns compared to those
MizBee
A Browser for Comparative Genomics Data

• data from the field of comparative genomics
• series of interviews with two biologists
• validate, analyze, and communicate computational results
high level biology questions

**evolution:** How long ago did two species share a common ancestor?

**function:** Which segment of the genome is responsible for a specific function in the cell?

low level data-centric questions

Are the paired features within a block contiguous?
Which chromosomes share conserved blocks?
Are similarity scores alike within a block?
difficult to answer **multiple questions** across a **range of scales** using computational algorithms alone

visually encode conservation relationships at different scales to validate, analyze, and communicate results
design decisions

- represent chromosomes as segments
- encode conservation
design decisions

- represent chromosomes
- encode conservation
design decisions

- represent chromosomes
- encode conservation
- layout
design decisions

- represent chromosomes as segments
- encode conservation
- layout
- encode similarity
stickleback

pufferfish
“The first time I saw my data in [MizBee] I was totally disappointed. The data was very noisy, and there were many small blocks that went to different chromosomes.”

“Honestly, I don't know. I don't think I would ever have gotten here. The noise was very hard see in the scatter plots while [MizBee] is much more unforgiving.”

*Genome-wide Synteny Through Highly Sensitive Sequence Alignment: Satsuma*  
M. Grabherr et al., Bioinformatics 2010.
reflections . . .
• defined process helps focus development

• fast, nimble prototyping is essential

• refine methodology for visualization design
  • particularly *abstract* and *validate* stages

*An Insight-based Methodology for Evaluating Bioinformatics Visualizations.*
*P. Saraiya et al., IEEE TVCG 2005.*
questions?

www.bangwong.com
www.iic.harvard.edu/~miriah

acknowledgements

Nils Gelhenborg

Data Visualization Initiative at the Broad Institute:
www.broadinstitute.org/vis

Regev lab: Michelle Chan, Courtney French, Jay Konieczka, Jenna Pfiffner, Aviv Regev, and Dawn Thompson

DePace lab: Angela DePace, Tara Martin, Sarah Peter, Max Staller, Zeba Wunderlich

Manfred Grabherr and Li-Jun Ma

National Science Foundation under Grant 0937060 to the Computing Research Association for the CIFellows Project